Response to O'Grady et al.: The potential and peril of the supertree approach

KIM VAN DER LINDE and DAVID HOULE

Insect Syst.Evol.

van der Linde, K. and Houle, D.: Response to O'Grady et al.: The potential and peril of the supertree approach. *Insect Syst. Evol.* 39: 281-286. Copenhagen, October 2008. ISSN1399-560X.

Corresponding author: Kim van der Linde, Department of Biological Science, Florida State University, Tallahassee, FL 32306-4295, U.S.A.; telephone (850) 645-8521, fax (850) 645-8447, email: kim@kimvdlinde.com.

Introduction

We welcome the critical response by O'Grady et al. (2008) to our article (van der Linde & Houle 2008), as it raises several interesting general points about supertrees and points out areas where details of our analysis were insufficiently clear. On the other hand, O'Grady et al. also misrepresent what we have done and what we conclude on the basis of our review.

O'Grady et al. make three broad points. First, they point out potential problems with the supertree approach that are, in practice, difficult to eliminate. Second they detail areas of our supertree analysis that they find deficient. We argue below that these rest either on their misunderstanding our methods (some of which results from our lack of clarity) or on deficiencies in the available data. Third, they object to the use of our supertree as evidence for taxonomic revision of the genus *Drosophila.* The taxonomic status of *Drosophila* is not the subject of our paper, although our results are certainly relevant to that discussion. We discuss each of these three broad points in turn.

The goal of our original paper was to provide as detailed a phylogenetic hypothesis for as wide a selection of taxa in the subfamily Drosophilinae as the data will support, by drawing on the extensive but scattered literature on drosophilid phylogeny. Knowledge of the genetics, genome, and biology of *Drosophila melanogaster* is among the best for any metazoan, and knowledge of many other species is also increasing rapidly, as exemplified by the 12-genome project (*Drosophila* 12 Genomes Consortium 2007). Many other species currently assigned to the genus *Drosophila* have served as models in evolutionary biology. This background information presents a huge opportunity for comparative biology in the larger clade. The lack of a comprehensive, well-supported phylogeny of the clade containing these species is a major impediment to exploiting this opportunity.

Given the potential scientific benefits of comparative analyses in the Drosophilidae, the lack of a comprehensive phylogeny for this group is a scandal. The symptoms of this deplorable situation are described in our original article. Although many fine studies of clades within the currently accepted subgenera of *Drosophila* are available, the coverage of the closely related genera is quite poor. Authors have noted strong evidence that the genus *Drosophila* is paraphyletic for over 30 years, yet the coverage of taxa not currently assigned to the genus *Drosophila* is still quite scant. Few studies use more than a handful of genes, and none employs sampling of taxa that could be considered adequate for phylogenetic questions at the family or subfamily level.

We therefore see the following contradictory situation. The opportunities for comparative biology in the Drosophilidae are very great, yet the data with which to construct a phylogenetic hypothesis are relatively weak. This situation is the important background to our supertree analysis and review.

Supertree analyses

O'Grady et al. (2008) capably outline the potential difficulties that may be encountered in assembling a supertree. We are in complete agreement with these points, and we cited literature covering the same issues in our paper. On the other hand, throughout their critique, O'Grady et al. imply that a supertree analysis is a simple matter of following a set of clearly defined best practices. Unfortunately the actual data available for most supertree analyses are far from ideal, and one is forced to balance taxonomic coverage, independence of data, and objective weighting criteria. These sorts of situations inevitably introduce a subjective element into supertree analyses. Many of O'Grady et al.'s (2008) criticisms are of precisely these inevitably subjective decisions.

Specific criticisms of our input data

O'Grady et al. (2008) criticize our selection of data in the following areas.

Statement of criteria for the choice of input trees

O'Grady et al. (2008) are right that we did not provide a sufficiently explicit statement about the selection of the input trees. We endeavoured to locate all the literature on drosophilid phylogenies and searched Web of Science, Google Scholar, TaxoDros (Bächli 1999-2008), and the literature lists of phylogenetic articles for relevant publications. For each publication, we determined the independent sources of data (e.g., genes, morphology, chromosomes; cf. Bininda-Emonds et al. 2004) and selected the most comprehensive tree for our analysis. If several trees were equally comprehensive in taxa covered, we favoured trees with an adequate nucleotide substitution model over trees that lacked such an approach, as the variation in nucleotide content varies dramatically within the family Drosophilidae (Anderson et al. 1993; Moriyama & Hartl 1993; Rodriguez-Trelles et al. 1999; Tarrio et al. 2000, 2001; Powell et al. 2003; Tamura et al. 2004; Clark et al. 2007; Heger & Ponting 2007). We also tried to minimize the overlap between articles using the same criteria (e.g., Kopp & True 2002 and Kopp 2006) but accepted some overlap in order to maximize the total species coverage.

The charitable reader will note that we raised data-selection issues in our original paper and presented some specific examples of our decision-making process. Furthermore the full list of trees we used and their weights is given in our paper.

Our designation of outgroups

O'Grady et al. (2008) claim that the composite outgroup used in our analysis included members of the family Drosophilidae. It did not. As explained in our article, our outgroup was a composite made up of data from only the non-drosophilids in our sample of phylogenies. O'Grady et al. (2008) also criticize us for using a composite outgroup at all, but this practice is standard in supertree analyses, where it is known as the semirooted MR-outgroup approach (Bininda-Emonds et al. 2004).

Not using input trees preferred by the original authors

O'Grady et al. (2008) claim that, when multiple trees are presented in one study, the choice of the trees preferred by the authors of the original paper should be preferred, and they cite two papers in support of this contention: Gatesy et al. (2002) and Bininda-Emonds et al. (2004). In fact, Gatesy et al. (2002) did not address this issue, whereas Bininda-Emonds et al. (2004) advocate a much more elaborate decision scheme than O'Grady et al. (2008) state. If we limit the suggested protocol to the section dealing with multiple trees within a single study, Bininda-Emonds et al. give the following sequence to determine which trees to include. The first step is to determine all independent data sources, such as single genes, and unique combinations of these independent data sources, as well as non-overlapping taxon sets for a single data source. Trees based on each of these data sources can be included. In the case of non-independent source trees, in other words, trees based on the same underlying data and same or overlapping taxon sets, they suggest using the most comprehensive tree, and if such a tree is not available, the tree explicitly preferred by the authors, and if that is not available, the consensus of the non-independent trees. Finally, if all else fails, they suggest constructing a mini-supertree and using it as a source tree in the supertree analysis.

Our approach closely mimics the suggested protocol of Bininda-Emonds et al. (2004). We first determined the independent data sources within a study and then selected the most comprehensive tree of those available for each independent data source. Contrary to the suggested protocol of Bin-

inda-Emonds et al. (2004), we included trees based on unique combinations of multiple independent data sources, such as the total-evidence tree, only if we did not include trees based on the various independent data sources, because we feel that including both is effectively pseudoreplication of the data, despite arguments advocating 'signal enhancement' (*sensu* de Queiroz et al. 1995).

Use of unpublished trees

O'Grady et al. are correct that we used two not-yet-peer-reviewed trees (van der Linde et al. in press). As noted by Bininda-Emonds et al. (2004), the use of unpublished trees can decrease accountability. On the other hand, omitting them will limit the taxonomic coverage of the analysis if relevant unpublished data are known to the authors. We have therefore provided O'Grady et al. with the unpublished manuscript so that they can check the validity of our analysis. We also obtained two trees reported on by Katoh et al. (2007) about which full details had not been included in the original paper.

Weighting of input trees

O'Grady et al. (2008) suggest that we used an inappropriate weighting scheme for our input trees and that the weighting scheme should have been based on measures of nodal support. We did not use this method because many of the source trees did not include such support measures (Bininda-Emonds & Sanderson 2001), whereas support measures obtained by different methods (e.g., posterior probabilities, Bremer support, bootstrap values) are difficult to compare (Cummings et al. 2003; Douady et al. 2003; Erixon et al. 2003; Svennblad et al. 2006). Limiting source trees to those with the same support-measure estimates would have greatly compromised the taxonomic coverage of our analysis and is contrary to common practice. Generally, weighting schemes are arbitrary based on, for example, large as opposed to small sets of data (Liu et al. 2001) or omitted altogether. Burleigh et al. (2006) suggest the use of alternative weighting schemes based on the number of variable characters or the total length of sequence, but they did not investigate the effect of those weighting schemes relative to alternatives. Those schemes favor base-pair counts over independence of sources. Our weighting scheme, to the contrary, favors the number of independent data sources. For example, a study that used a single long sequence would not be weighted as heavily as one that used several independent shorter sequences. In the absence of formal studies examining the various weighting alternatives, the choice of either weighting scheme is subjective and open to criticism.

We clearly stated in our original article that the weights were based on the numbers of genes used for the tree, with three exceptions, to all noted by us in the original paper. The first was to substitute the average number of genes per species in those studies where the average was considerably lower than the total number of genes used. The three such exceptions are all listed in Table 2. In addition, we assigned a weight of 5 to the 12-genome tree from *Drosophila* 12 Genomes Consortium (2007), even though it is based on sequence from over 300 genes. O'Grady et al. provide contradictory criticisms on this point. They imply, on the one hand, that it should be weighted more heavily because of the large number of genes and, on the other, that it should be less heavily weighted because of the small number of taxa. Clearly, however, increasing the weighting factor for this source tree would not have changed the outcome, as the 12-genome tree is actually congruent with the resulting supertree, as we stated in our article. A final exception to our weighting scheme was to reduce the weight of the study by Oliviera et al. (2005) to compensate for the large overlap of this study with others already in our sample. The last two weighting decisions are subjective but do have clear rationales.

Data analysis

O'Grady et al. (2008) criticize our analyses for sampling tree space too sparsely and for not estimating support for the nodes in our tree. The reanalysis of our weighted data by O'Grady et al. (2008) resulted in a nearly identical tree; the only difference from our tree was in the *tripunctata* clade, where a solution one step more parsimonious was obtained. Their unweighted tree was substantially less well resolved, as was ours, the very reason why the use of weighted supertrees is recommended (Bininda-Emonds & Sanderson 2001).

Sampling tree space

A major criticism by O'Grady et al. (2008) is that we might have searched tree space insufficiently and that many more equally parsimonious or even

more parsimonious trees could have been found. This issue besets all heuristic searches that include large numbers of taxa; they are not guaranteed to find the best solution for a variety of reasons, including arrest to a suboptimal local minimum.

The analysis of this set of data produces a large number of equally parsimonious trees, a result essentially caused by the multiplicative effect of several small local clades that are not well resolved. Each such clade generates a finite number of equally parsimonious solutions, but for a tree that contains more than one such poorly resolved clade, the number of equally parsimonious solutions is the product of the number of solutions for each such clade.

To determine whether we had missed additional, more parsimonious trees, we analyzed all major clades independently, without limiting the Max Tree variable. For the clades in question, we subjected them to searches using closest and random addition, as well as using a starting tree. The best results were identical to our previous results, with the exception of the *immigrans-tripunctata* clade, where we recovered the same one-step-shorter solution found by O'Grady et al. (2008). Each of the section analyses resulted in a limited number of equally parsimonious trees (*Sophophora*: 158 trees; Hawaiian Drosophilidae: 35 trees; *virilis-repleta* radiation: 720 trees; *immigrans-tripunctata* radiation: 7 trees). Multiplying these numbers produces a prediction of 27,871,200 equally parsimonious solutions. All but one of the polytomies were caused by either a single or a few unstable species. When those species were removed, analysis of the entire combined set of data yielded only 16 equally parsimonious trees, with exactly the same topology as our original analysis except for the absence of the removed species.

Another major criticism by O'Grady et al. (2008) is that we did not obtain statistical support measures for our tree, for example by source-tree bootstrapping. We agree that estimating support is very desirable. We did not do so because the nature of the input data would cause such measures to be biased. Bootstrapping of sequence data by nucleotide gives sensible measures of support because of the exchangeability of each site in a sequence. That is, no site is expected, a priori, to be more informative than any other site. This exchangeability property is seriously violated in the case of the source trees in our analysis. Studies in our sample differ in the number of genes used (and in whether sequence was used) but, more importantly, in taxonomic coverage. Studies in our sample used as few as 4 taxa and as many as 165. Resampling at the level of source trees can therefore only result in coverage of taxa that is less balanced than that in our original selection of trees. As shown in our original Figure 2 and in our accompanying text, taxa that are only included in sparsely sampled or small numbers of trees will necessarily have a more uncertain position than that estimated from the full data set. The probability that any particular tree will not be represented in a particular bootstrap sample is $1/e = 0.368$. The minimum number of trees a taxon had to be in before we included it in our analysis was only three. Each taxon appearing in just three trees will be missing in 5% of all bootstrap samples, present once in 9%, and present twice in 15%.

Supertrees as guides for taxonomy

We agree with O'Grady et al. that taxonomic revisions should be based on solid evidence. As a result, all evidence present in the studies covering a single group should be considered and the results of a single study not permitted to dictate the outcome of such an analysis. Therefore, proposals to revise a taxon are always based on combining the data of many studies. Formal supertrees are one tool that can be used for this purpose.

Conclusion

The primary goal of our supertree analysis of the family Drosophilidae was to provide a phylogenetic hypothesis for the group to further our ability to do comparative analyses. Of particular concern to us was to maximize the number of taxa that were covered by our analysis while still preserving adequate data from which to draw conclusions. These twin goals led us to many compromises, many of which inevitably have subjective components. Clearly, other authors might have made such decisions differently. Overall, we believe that the decisions we have made were sound. We will be happy to furnish the data on which our analysis is based to any researcher who wishes to investigate the effects of these decisions on the outcome (as we have already done with O'Grady et al.) or to perform more extensive analyses as more data become available.

In summary, we agree with most of the general

points about supertrees raised by O'Grady et al., but most of their criticisms of our study are based on misunderstanding or misrepresentation of our article. As we have shown, our analysis was relatively conservative (e.g., omitting total-evidence trees when we used the individual gene trees of a publication, contra Bininda-Emondset al. 2004, see above), and the resulting tree was probably very close to the most parsimonious tree. We have also shown that our analyses are very unlikely to have missed a substantially different, more parsimonious tree. This is also reflected in our reanalysis of the data – omitting the two trees by van der Linde et al. (in press) and including the maximum-parsimony tree recently published by O'Grady & DeSalle (2008) – which resulted in tree largely congruent with our tree. To conclude, the tree presented in our paper forms a solid hypothesis for the phylogenetic relationships within the genus *Drosophila* and its included genera.

References

Anderson, C.L., Carew, E.A. & Powell, J.R. (1993) Evolution of the Adh locus in the *Drosophila willistoni* group: the loss of an intron, and shift in codon usage. *Molecular Biology and Evolution* 10: 605–618.

Bächli, G. (1999-2008) TaxoDros: The Database on Taxonomy of Drosophilidae. Available at: http://taxodros.unizh.ch/

Bininda-Emonds, O.R.P., Beck, R.M.D. & Purvis, A. (2005) Getting to the roots of matrix representation. *Systematic Biology* 54: 668-U7.

Bininda-Emonds, O.R.P., Jones, K.E., Price, S.A., Cardillo, M., Grenyes, R. & Purvis, A. (2004) Garbage in, garbage out: data issues in supertree construction. Pp. 267-277 *in* O.R.P. Bininda-Emonds: *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*. Kluwer Academic Publishers, Boston, MA.

Bininda-Emonds, O.R.P. & Sanderson, M.J. (2001) Assessment of the accuracy of matrix representation with parsimony analysis supertree construction. *Systematic Biology* 50: 565–579.

Burleigh, J.G., Driskell, A.C. & Sanderson, M.J. (2006) Supertree bootstrapping methods for assessing phylogenetic variation among genes in genome-scale data sets. *Systematic Biology* 55: 426-440.

Clark, A.G., Eisen, M.B., Smith, D.R., Bergman, C.M., Oliver, B., Markow, T.A., Kaufman, T.C., Kellis, M., Gelbart, W., Iyer, V.N. (2007) Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* 450: 203–218.

Cummings, M.P., Handley, S.A., Myers, D.S., Reed, D.L., Rokas, A. & Winka, K. (2003) Comparing bootstrap and posterior probability values in the four-taxon case. *Systematic Biology* 52: 477-487.

de Queiroz, A., Donoghue, M.J. & Kim, J. (1995) Separate versus combined analysis of phylogenetic evidence. *Annual Review of Ecology and Systematics* 26: 657-681.

Douady, C.J., Delsuc, F., Boucher, Y., Doolittle, W.F. & Douzery, E.J.P. (2003) Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. *Molecular Biology and Evolution* 20: 248-254.

Drosophila 12 Genomes Consortium (2007) Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* 450: 203–218.

Erixon, P., Svennblad, B., Britton, T. & Oxelman, B. (2003) Reliability of Bayesian posterior probabilities and bootstrap frequencies in phylogenetics. *Systematic Biology* 52: 665-673.

Gatesy, J., Matthee, C., DeSalle, R. & Hayashi, C. (2002) Resolution of a supertree/supermatrix paradox. *Systematic Biology* 51: 652–664.

Heger, A. & Ponting, C. P. (2007) Variable strength of translational selection among 12 *Drosophila* species. *Genetics* 177: 1337–1348.

Katoh, T., Nakaya, D., Tamura, K. & Aotsuka, T. (2007) Phylogeny of the *Drosophila immigrans* species group (Diptera: Drosophilidae) based on Adh and Gpdh sequences. *Zoological Science* 24: 913–921.

Kopp, A. (2006) Basal relationships in the *Drosophila melanogaster* species group. *Molecular Phylogenetics and Evolution* 39: 787–798.

Kopp, A. & True, J.R. (2002) Phylogeny of the oriental *Drosophila melanogaster* species group: a multilocus reconstruction. *Systematic Biology* 51: 786–805.

Liu, F.G.R., Miyamoto, M.M., Freire, N.P., Ong, P.Q., Tennant, M.R., Young, T.S. & Gugel, K.F. (2001) Molecular and morphological supertrees for Eutherian (placental) mammals. *Science* 291: 1786–1789.

Moriyama, E.N. & Hartl, D.L. (1993) Codon usage bias and base composition of nuclear genes in *Drosophila*. *Genetics* 134: 847–858.

O'Grady, P.M. & DeSalle, R. (2008) Out of Hawaii: the origin and biogeography of the genus *Scaptomyza* (Diptera: Drosophilidae). *Biology Letters* 4: 195-199.

O'Grady, P.M., Lapoint, R.T. & Bennett, G.M. (2008) The potential and peril of the supertree approach: A response to van der Linde and Houle. *Insect Systematics and Evolution* 39: 269-280.

Oliveira, D., O'Grady, P.M., Etges, W.J., Heed, W.B. & DeSalle, R. (2005) Molecular systematics and geographical distribution of the *Drosophila longicornis* species complex (Diptera : Drosophilidae). *Zootaxa*: 1–32.

Powell, J.R., Sezzi, E., Moriyama, E.N., Gleason, J.M. & Caccone, A. (2003) Analysis of a shift in codon usage in *Drosophila*. *Journal of Molecular Evolution* 57: S214–S225.

Rodriguez-Trelles, F., Tarrio, R. & Ayala, F.J. (1999) Switch in codon bias and increased rates of amino acid substitution in the *Drosophila saltans* species group. *Genetics* 153: 339–350.

Svennblad, B., Erixon, P., Oxelman, B. & Britton, T. (2006) Fundamental differences between the methods of maximum likelihood and maximum posterior probability in phylogenetics. *Systematic Biology* 55: 116-121.

Tamura, K., Subramanian, S. & Kumar, S. (2004) Temporal patterns of fruit fly (*Drosophila*) evolution revealed by mutation clocks. *Molecular Biology and Evolution* 21: 36–44.

Tarrio, R., Rodriguez-Trelles, F. & Ayala, F.J. (2000) Tree rooting with outgroups when they differ in their nucleotide composition from the ingroup: the *Drosophila saltans* and *willistoni* groups, a case study. *Molecular Phylogenetics and Evolution* 16: 344–349.

Tarrio, R., Rodriguez-Trelles, F. & Ayala, F.J. (2001) Shared nucleotide composition biases among species and their impact on phylogenetic reconstructions of the Drosophilidae. *Molecular Biology and Evolution* 18: 1464–1473.

van der Linde, K., Bächli, G., Toda, M. J., Zhang, W.-X., Katoh, T., Hu, Y.-G. & Spicer, G.S. (in press) Resolving the paraphyletic status of the genus *Drosophila* while preserving the name of *Drosophila melanogaster*. *Journal of Zoological Systematics and Evolutionary Research*.

van der Linde, K. & Houle, D. (2008) A supertree analysis and literature review of the genus *Drosophila* and related genera. *Insect Systematics and Evolution* 39: xxx-xxx.

Accepted for publication September 2008